Novelty Assessment Report

Paper: Stable Video Infinity: Infinite-Length Video Generation with Error Recycling

PDF URL: https://openreview.net/pdf?id=X96Ei9n34a

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

We propose **Stable Video Infinity (SVI)** that can generate non-looping, ultra-long videos with stable visual quality, while supporting per-clip prompt control and multi-modal conditioning. While existing long-video methods attempt to **mitigate accumulated errors** via handcrafted anti-drifting (e.g., modified noise scheduler, frame anchoring), they remain limited to single-prompt extrapolation, producing homogeneous scenes with repetitive motions. We identify that the fundamental challenge extends beyond error accumulation to a critical discrepancy between the training assumption (seeing clean data) and the test-time autoregressive reality (conditioning on self-generated, error-prone outputs). To bridge this hypothesis gap, SVI incorporates **Error-Recycling Fine-Tuning**, a new type of efficient training that recycles the Diffusion Transformer (DiT)'s self-generated errors into supervisory prompts, thereby encouraging DiT to **actively identify and correct its own errors**. This is achieved by injecting, collecting, and banking errors through closed-loop recycling, autoregressively learning from error-injected feedback. Specifically, we (i) inject historical errors made by DiT to intervene on clean inputs, simulating error-accumulated trajectories in flow matching; (ii) efficiently approximate predictions with one-step bidirectional integration and calculate errors with residuals; (iii) dynamically bank errors into replay memory across discretized timesteps, which are resampled for new input. SVI is able to scale videos from seconds to infinite durations with no additional inference cost, while remaining compatible with diverse conditions (e.g., audio, skeleton, and text streams). We evaluate SVI on three benchmarks, including consistent, creative, and conditional settings, thoroughly verifying its versatility and state-of-the-art role. Project page

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Infinite-Length Video Generation with Stable Quality**

A total of **50 papers** were analyzed and organized into a taxonomy with **22 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Autoregressive and Causal Generation Frameworks**
- **Diffusion-Based Temporal Modeling**
- **Hybrid Autoregressive-Diffusion Architectures**
- **Error Mitigation and Consistency Mechanisms**
- **Structured Scene and World Modeling**
- **Planning and Hierarchical Decomposition**
- **Compositional and Object-Centric Approaches**
- **Specialized Application Domains**
- **Evaluation and Quality Assessment**
- **Auxiliary Enhancement Techniques**

Complete Taxonomy Tree

- Infinite-Length Video Generation with Stable Quality Survey Taxonomy
- Autoregressive and Causal Generation Frameworks
  - Real-Time Streaming Generation (3 papers)
  - [2] MotionStream: Real-Time Video Generation with Interactive Motion Controls (Shin, 2025) View paper
  - [8] Longlive: Real-time interactive long video generation (Yang, 2025) View paper
  - [47] Rolling Forcing: Autoregressive Long Video Diffusion in Real Time (Liu, 2025) View paper
  - Long-Horizon Autoregressive Synthesis (4 papers)
  - [10] Macro-from-micro planning for high-quality and parallelized autoregressive long video generation (Chen Ya-bo, 2025) View paper
  - [13] DeepVerse: 4D Autoregressive Video Generation as a World Model (Chen Junyi, 2025) View paper
  - [39] BAgger: Backwards Aggregation for Mitigating Drift in Autoregressive Video Diffusion Models (Ryan Po, 2025) View paper
  - [43] Markov Decision Process for Video Generation (Vladyslav Yushchenko, 2019) View paper
  - Audio-Driven Avatar Generation (4 papers)
  - [6] Live Avatar: Streaming Real-time Audio-Driven Avatar Generation with Infinite Length (Yubo Huang, 2025) View paper
  - [26] JoyAvatar: Real-time and Infinite Audio-Driven Avatar Generation with Autoregressive Diffusion (Chaochao Li, 2025) View paper
  - [29] StableAvatar: Infinite-Length Audio-Driven Avatar Video Generation (Tu, 2025) View paper
  - [30] MagicInfinite: Generating Infinite Talking Videos with Your Words and Voice (Yi, 2025) View paper
- Diffusion-Based Temporal Modeling
  - Short-Term Memory and Conditional Attention (3 papers)
  - [4] Streamingt2v: Consistent, dynamic, and extendable long video generation from text (Roberto Henschel, 2025) View paper

  - [5] LaVie: High-Quality Video Generation with Cascaded Latent Diffusion Models (Yaohui Wang, 2023) View paper
  - [19] MALT Diffusion: Memory-Augmented Latent Transformers for Any-Length Video Generation (Yu, 2025) View paper
  - Cascaded and Hierarchical Diffusion (3 papers)
  - [7] Vidgen: Long-Form Text-to-Video Generation with Temporal, Narrative and Visual Consistency for High Quality Story-Visualisation Tasks (Ram Selvaraj, 2024) View paper
  - [20] SkyReels-Audio: Omni Audio-Conditioned Talking Portraits in Video Diffusion Transformers (Fei, 2025) View paper
  - [23] Longvie: Multimodal-guided controllable ultra-long video generation (Gao Jian-xiong, 2025) View paper
  - Motion and Trajectory Control (3 papers)
  - [1] MimicMotion: High-Quality Human Motion Video Generation with Confidence-aware Pose Guidance (Zhang Yu-ang, 2024) View paper
  - [15] Tora: Trajectory-oriented Diffusion Transformer for Video Generation (Zhenghao Zhang, 2024) View paper
  - [24] Frame In-N-Out: Unbounded Controllable Image-to-Video Generation (Wang Boyang, 2025) View paper
  - Training-Free Extension Techniques (2 papers)
  - [16] FreeLong: Training-Free Long Video Generation with SpectralBlend Temporal Attention (Yuanzhi Liang, 2024) View paper
  - [37] Brick-Diffusion: Generating Long Videos with Brick-to-Wall Denoising (Yunlong Yuan, 2025) View paper
- Hybrid Autoregressive-Diffusion Architectures
  - Coarse-to-Fine Guidance Integration (2 papers)
  - [21] ARLON: Boosting Diffusion Transformers with Autoregressive Models for Long Video Generation (Li, 2024) View paper
  - [45] TempoMaster: Efficient Long Video Generation via Next-Frame-Rate Prediction (Yukuo Ma, 2025) View paper
  - Block and Semi-Autoregressive Diffusion (2 papers)
  - [33] Self-Forcing++: Towards Minute-Scale High-Quality Video Generation (Cui, 2025) View paper
  - [38] BlockVid: Block Diffusion for High-Quality and Consistent Minute-Long Video Generation (Zeyu Zhang, 2025) View paper
  - Distillation and Distribution Matching (1 papers)
  - [25] Reward Forcing: Efficient Streaming Video Generation with Rewarded Distribution Matching Distillation (Yunhong Lu, 2025) View paper
- Error Mitigation and Consistency Mechanisms
  - Error Recycling and Self-Correction ★ (1 papers)
  - [0] Stable Video Infinity: Infinite-Length Video Generation with Error Recycling (Anon et al., 2026) View paper
  - Adaptive Caching and Inference Optimization (1 papers)
  - [22] Adaptive Caching for Faster Video Generation with Diffusion Transformers (Kahatapitiya, 2024) View paper
- Structured Scene and World Modeling
  - 4D and Geometric World Models (3 papers)
  - [12] Generating long videos of dynamic scenes (Brooks, 2022) View paper
  - [31] Make-it-4d: Synthesizing a consistent long-term dynamic scene video from a single image (Liao Shen, 2023) View paper
  - [34] LongScape: Advancing Long-Horizon Embodied World Models with Context-Aware MoE (Shang Yu, 2025) View paper
  - 3D Scene Generation and Camera Control (5 papers)
  - [11] Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models (Lu, 2025) View paper
  - [27] DreamJourney: Perpetual View Generation with Video Diffusion Models (Pan Bo, 2025) View paper
  - [28] Persistent nature: A generative model of unbounded 3d worlds (Lucy Chai, 2023) View paper
  - [49] BulletTime: Decoupled Control of Time and Camera Pose for Video Generation (Yiming Wang, 2025) View paper
  - [50] Captain Safari: A World Engine (Yu-Cheng Chou, 2025) View paper
  - Driving Scene Simulation (2 papers)
  - [17] DriveGen3D: Boosting Feed-Forward Driving Scene Generation with Efficient Video Diffusion (Wang Wei-jie, 2025) View paper
  - [36] STAGE: A Stream-Centric Generative World Model for Long-Horizon Driving-Scene Simulation (Wang Jiamin, 2025) View paper
- Planning and Hierarchical Decomposition (3 papers)
  - [3] SkyReels-V2: Infinite-length Film Generative Model (Chen Gui-bin, 2025) View paper
  - [14] Video is worth a thousand images: Exploring the latest trends in long video generation (Faraz Waseem, 2025) View paper
  - [41] WorldWeaver: Generating Long-Horizon Video Worlds via Rich Perception (Liu Zhi-heng, 2025) View paper
- Compositional and Object-Centric Approaches (3 papers)
  - [32] Compositional Video Synthesis by Temporal Object-Centric Learning (Akan, 2025) View paper
  - [42] AI Powered High Quality Text to Video Generation with Enhanced Temporal Consistency (Patel, 2025) View paper
  - [46] Animate Any Character in Any World (Yitong Wang, 2025) View paper
- Specialized Application Domains
  - Robotic Manipulation and Embodied AI (1 papers)
  - [18] RoboEnvision: A Long-Horizon Video Generation Model for Multi-Task Robot Manipulation (Yang Liu-di, 2025) View paper
  - Immersive and 360-Degree Content (2 papers)
  - [40] AI and Generative Models in 360-Degree Video Creation: Building the Future of Virtual Realities (Nicolay Anderson Christian, 2025) View paper
  - [48] Unboxed: Geometrically and Temporally Consistent Video Outpainting (Zhongrui Yu, 2025) View paper
  - Interactive and Controllable Character Animation (1 papers)
  - [44] TalkVerse: Democratizing Minute-Long Audio-Driven Video Generation (Zhenzhi Wang, 2025) View paper
- Evaluation and Quality Assessment (1 papers)
  - [35] LMVQ: Label-free Metric-learning for General AI-generated Video Quality Assessment (Zhichao Zhang, 2025) View paper
- Auxiliary Enhancement Techniques (1 papers)
  - [9] 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors (Siyu Huang, 2024) View paper

Narrative

Core task: infinite-length video generation with stable quality. The field addresses the challenge of producing arbitrarily long video sequences without accumulating visual artifacts or degrading coherence over time. The taxonomy reveals a diverse landscape organized around ten major branches. Autoregressive and causal generation frameworks focus on sequential frame prediction, while diffusion-based temporal modeling leverages iterative denoising processes for high-quality synthesis. Hybrid autoregressive-diffusion architectures combine both paradigms to balance efficiency and fidelity. Error mitigation and consistency mechanisms tackle drift and quality decay through techniques like error recycling and self-correction. Structured scene and world modeling approaches build explicit representations of environments, whereas planning and hierarchical decomposition methods break generation into manageable subproblems. Compositional and object-centric approaches decompose scenes into reusable elements, and specialized application domains target specific use cases such as avatars or driving scenarios. Evaluation and quality assessment branches develop metrics for long-form coherence, while auxiliary enhancement techniques provide supporting tools like caching or reward shaping.

Several active lines of work highlight contrasting strategies for maintaining stability. Autoregressive methods like StreamingT2V[4] and FreeLong[16] emphasize efficient temporal extension through sliding windows and memory mechanisms, but face compounding error challenges. Diffusion-based approaches such as LaVie[5] and MotionStream[2] prioritize visual quality but require careful temporal conditioning. Stable Video Infinity[0] sits within the error mitigation and consistency mechanisms branch, specifically addressing error recycling and self-correction. Its emphasis on actively detecting and correcting accumulated errors distinguishes it from purely autoregressive methods like StreamingT2V[4], which rely on conditioning strategies, and from diffusion-heavy approaches like SkyReels[3], which focus on temporal coherence through architectural design. The work reflects a growing recognition that infinite-length generation demands explicit mechanisms to counteract drift, rather than relying solely on model capacity or temporal attention.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

Both subtopics address computational efficiency and quality maintenance in long video generation, but from different angles. Error Recycling and Self-Correction focuses on training models to detect and fix their own mistakes through feedback mechanisms, while Adaptive Caching and Inference Optimization emphasizes runtime efficiency through memory management and token optimization without explicit error correction loops.

Similarities: - Both aim to improve quality and stability in extended video generation - Both address error propagation challenges inherent in autoregressive or long-sequence generation - Both seek to optimize computational resources during inference

Differences: - Error Recycling trains models with error injection and closed-loop feedback for self-correction capabilities, while Adaptive Caching uses architectural optimizations like sliding windows and caching - Error Recycling modifies the training paradigm to build correction abilities, whereas Adaptive Caching focuses on inference-time efficiency without retraining - Error Recycling explicitly handles mistakes through identification and correction mechanisms, while Adaptive Caching prevents errors through better resource management and reduced propagation

Suggested Search Directions: - Hybrid approaches combining error correction feedback with adaptive caching mechanisms - Comparative studies on whether error prevention (caching) or error correction (recycling) is more effective for infinite-length generation - Methods that use cached states as checkpoints for error recovery in self-correcting models

Sibling Subtopics

- **Adaptive Caching and Inference Optimization** (leaves: 1, papers: 1)
- Scope: Techniques using adaptive caching, sliding windows, or dynamic token management to reduce computational cost and error propagation.
- Exclude: Error recycling methods belong to Error Recycling and Self-Correction; general autoregressive methods belong to Autoregressive and Causal Generation Frameworks.

Contributions Analysis

Overall novelty summary. The paper proposes Stable Video Infinity (SVI), a system for generating ultra-long, non-looping videos with per-clip prompt control through Error-Recycling Fine-Tuning. Within the taxonomy, it occupies the 'Error Recycling and Self-Correction' leaf under 'Error Mitigation and Consistency Mechanisms'. Notably, this leaf contains only the original paper itself—no sibling papers exist in this specific category. This positioning suggests the work addresses a relatively sparse research direction focused explicitly on training models to identify and correct their own errors through closed-loop feedback.

The taxonomy reveals that error mitigation strategies are distributed across multiple branches. The parent branch 'Error Mitigation and Consistency Mechanisms' also includes 'Adaptive Caching and Inference Optimization', which addresses error propagation through computational techniques rather than self-correction. Neighboring branches like 'Autoregressive and Causal Generation Frameworks' (containing StreamingT2V and FreeLong) tackle error accumulation through architectural choices and conditioning strategies, while 'Diffusion-Based Temporal Modeling' methods prioritize temporal coherence through attention mechanisms. SVI's approach diverges by explicitly training on self-generated errors rather than relying on inference-time modifications or architectural constraints.

Among the three contributions analyzed, the literature search examined 27 candidates total. The core SVI system and Error-Recycling Fine-Tuning method each examined 10 candidates with zero refutable matches, suggesting these specific mechanisms appear novel within the limited search scope. However, the formalization of training-test hypothesis gap examined 7 candidates and found 5 refutable matches, indicating substantial prior work on exposure bias and distribution shift in autoregressive generation. The statistics reflect a focused semantic search rather than exhaustive coverage, so contributions appearing novel here may have relevant precedents outside the top-27 candidates examined.

Based on the limited search scope of 27 semantically similar papers, the error recycling mechanism appears to occupy underexplored territory within the taxonomy's sparse 'Error Recycling and Self-Correction' leaf. The formalization of hypothesis gap, however, connects to established literature on exposure bias. The analysis captures relationships within top-ranked semantic matches but does not claim comprehensive coverage of all relevant prior work in autoregressive video generation or error mitigation strategies.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Stable Video Infinity for infinite-length video generation

Description: The authors introduce Stable Video Infinity, a system that generates arbitrarily long videos without looping artifacts while maintaining stable quality. It supports per-clip prompt control and diverse multi-modal conditions such as audio and skeleton inputs.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Frame-Level Captions for Long Video Generation with Complex Multi Scenes

URL: View paper

Brief Assessment

Frame-Level Captions[54] focuses on frame-level text annotations and semantic confusion in multi-scene videos, not on error-recycling mechanisms or infinite-length generation systems. The candidate addresses dataset annotation methodology rather than the error correction framework proposed in the original paper.

2. Dreampose: Fashion video synthesis with stable diffusion

URL: View paper

Brief Assessment

DreamPose Fashion[57] focuses on fashion image-to-video synthesis from still images using pose conditioning, not on infinite-length video generation or error-recycling mechanisms for ultra-long videos.

3. Longlive: Real-time interactive long video generation

URL: View paper

Brief Assessment

LongLive[8] focuses on real-time interactive generation with streaming prompt inputs and kv-recache mechanisms for prompt switching, while the original paper addresses error accumulation through error-recycling fine-tuning. These represent distinct technical approaches to long video generation challenges.

4. Structure and content-guided video synthesis with diffusion models

URL: View paper

Brief Assessment

Structure Content Guided[51] focuses on structure-preserving video editing with depth conditioning and does not address infinite-length generation or error accumulation challenges that are central to the original paper's contribution.

5. Storydiffusion: Consistent self-attention for long-range image and video generation

URL: View paper

Brief Assessment

StoryDiffusion[56] focuses on generating consistent character images and short transition videos for storytelling, not infinite-length video generation with error correction mechanisms. The candidate addresses character consistency across image sequences, while the original tackles ultra-long video generation through error-recycling fine-tuning.

6. Make-your-video: Customized video generation using textual and structural guidance

URL: View paper

Brief Assessment

Make Your Video[55] focuses on customized video generation with structural guidance (e.g., depth maps) for precise control, using a two-stage learning scheme and causal attention masks for longer synthesis. It does not address infinite-length generation, error accumulation correction, or per-clip prompt control as core contributions.

7. Align your latents: High-resolution video synthesis with latent diffusion models

URL: View paper

Brief Assessment

Align Your Latents[53] focuses on high-resolution video synthesis using latent diffusion models with temporal alignment layers, but does not address infinite-length generation, error accumulation correction, or per-clip prompt control with multi-modal conditions like audio and skeleton inputs that are central to the original paper's contribution.

8. Text2story: Advancing video storytelling with text guidance

URL: View paper

Brief Assessment

Text2Story[58] focuses on long-form storytelling through prompt mixing and latent blending for scene/action transitions, not on infinite-length generation with error-recycling mechanisms or per-clip prompt control as in the original paper.

9. Self-Forcing++: Towards Minute-Scale High-Quality Video Generation

URL: View paper

Brief Assessment

Self-Forcing++[33] focuses on autoregressive video generation using distillation from short-horizon teachers and rolling KV cache mechanisms, whereas the original paper addresses error accumulation through error-recycling fine-tuning that injects and corrects self-generated errors. The technical approaches and underlying mechanisms differ fundamentally.

10. Dreampose: Fashion image-to-video synthesis via stable diffusion

URL: View paper

Brief Assessment

DreamPose[52] focuses on fashion image-to-video synthesis from a single image and pose sequence, not on infinite-length video generation or error-recycling mechanisms for ultra-long videos.

Contribution 2: Error-Recycling Fine-Tuning method

Description: A novel training approach that repurposes the model's own prediction errors as supervisory signals. This method enables the Diffusion Transformer to learn to identify and correct its mistakes through autoregressive error feedback, bridging the gap between error-free training and error-prone inference.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Towards dynamic modeling of visual-vestibular conflict detection

URL: View paper

Brief Assessment

Visual-Vestibular Conflict[70] focuses on drift-diffusion models for sensory conflict detection in human perception, not on training diffusion transformers for video generation using self-generated prediction errors as supervisory signals.

2. Rolling Forcing: Autoregressive Long Video Diffusion in Real Time

URL: View paper

Brief Assessment

Rolling Forcing[47] addresses error accumulation through rolling-window joint denoising and attention sink mechanisms, not through recycling self-generated errors as supervisory signals for training.

3. ETC: training-free diffusion models acceleration with Error-aware Trend Consistency

URL: View paper

Brief Assessment

ETC[68] focuses on training-free acceleration by reusing model outputs across timesteps to predict future denoising trends, not on training methods that use self-generated errors as supervisory signals. The candidate addresses inference-time trajectory consistency, while the original contribution concerns a novel training approach with autoregressive error feedback.

4. Self-guided diffusion models

URL: View paper

Brief Assessment

Self-guided Diffusion[64] focuses on replacing human annotations with self-supervised signals for guidance in diffusion models, not on using prediction errors as supervisory signals for training. The candidate addresses annotation-free guidance while the original addresses error accumulation in autoregressive video generation.

5. Rethinking Training Dynamics in Scale-wise Autoregressive Generation

URL: View paper

Brief Assessment

Scale-wise Autoregressive[69] addresses exposure bias in scale-wise visual generation through student-forcing during training, while the original paper focuses on video generation with error injection and recycling across temporal clips. The technical approaches and application domains differ fundamentally.

6. Your diffusion model is secretly a noise classifier and benefits from contrastive training

URL: View paper

Brief Assessment

Noise Classifier[67] focuses on improving diffusion model training through contrastive loss to distinguish noise levels, not on repurposing prediction errors as supervisory signals for autoregressive video generation.

7. Diffrect: Latent diffusion label rectification for semi-supervised medical image segmentation

URL: View paper

Brief Assessment

DiffRect[66] addresses pseudo-label quality in semi-supervised medical image segmentation through label rectification using diffusion models, not through recycling self-generated errors as supervisory signals for autoregressive error correction in video generation.

8. OViP: Online Vision-Language Preference Learning

URL: View paper

Brief Assessment

OViP[71] addresses hallucination in vision-language models through online preference learning with dynamically constructed contrastive data, while the original paper's error-recycling fine-tuning targets accumulated errors in autoregressive video generation by recycling the model's own prediction errors as supervisory signals for diffusion transformers.

9. Cm-gan: Stabilizing gan training with consistency models

URL: View paper

Brief Assessment

CM-GAN[65] focuses on stabilizing GAN training using consistency constraints from diffusion models, not on training diffusion models using self-generated errors as supervisory signals for autoregressive video generation.

10. One to Two, Two to All: Towards Multimodal Self-supervised Learning for Earth Observation

URL: View paper

Brief Assessment

Multimodal Earth Observation[72] focuses on self-supervised learning for Earth observation data using contrastive learning and denoising diffusion. It does not address using model prediction errors as supervisory signals for autoregressive video generation.

Contribution 3: Formalization of training-test hypothesis gap and error types

Description: The authors provide a systematic analysis identifying the fundamental discrepancy between training assumptions (clean data) and test-time reality (error-prone outputs). They formally define two error types: single-clip predictive error and cross-clip conditional error.

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Time-series generation by contrastive imitation

URL: View paper

Brief Assessment

Contrastive Imitation[62] addresses a training-test discrepancy in time-series generation (autoregressive models trained on clean data vs. conditioned on self-generated outputs), but focuses on compounding error in sequential generation rather than video-specific error propagation across clips. The original paper's formalization of single-clip predictive error and cross-clip conditional error is specific to video generation with spatial-temporal dynamics.

2. Rolling Forcing: Autoregressive Long Video Diffusion in Real Time

URL: View paper

Prior Art Analysis

Rolling Forcing[47] explicitly identifies and formalizes the training-test hypothesis gap in autoregressive video generation. The candidate paper states that 'During training, models are typically exposed to clean, error-free inputs; at test time, however, autoregressive generation conditions each new token (or frame) on previously generated outputs, which may contain errors' and describes this as a 'train-test gap known as exposure bias.' This demonstrates prior recognition of the fundamental discrepancy between training assumptions and test-time reality that the original paper claims to systematically analyze.

Evidence

Evidence 1 - **Rationale**: Both papers formalize the same concept that training assumes error-free historical trajectories while inference uses model predictions, identifying this as a fundamental gap. - **Original**: in training generative dit (fig. 1a), historical trajectories of flow matching are assumed to be error-free. however, this is easily broken in test since the model autoregressively uses previous generations with predictive errors - **Candidate**: In tf, the conditional distribution for theith frame at noise levelt j isp(x i tj |x <i 0), where all conditional history frames are the ground-truth clean frames from the training data. While in df, the conditional distribution isp(x i tj |x <i t≥0), where the history frames are the ground-truth ...

3. From Denoising to Refining: A Corrective Framework for Vision-Language Diffusion Model

URL: View paper

Prior Art Analysis

Denoising to Refining[61] demonstrates that the training-test hypothesis gap and error propagation in autoregressive generation were previously identified and formalized. The candidate paper explicitly identifies and formalizes the 'train-inference discrepancy' where models are 'trained exclusively on clean, ground-truth data but are required at inference to generate from their own noisy, intermediate outputs.' This directly parallels the original paper's claim about the discrepancy between training assumptions (clean data) and test-time reality (error-prone outputs). Both papers formalize how initial errors propagate and compound during generation, with the candidate describing this as 'error cascade driven by a training-inference discrepancy' and providing detailed analysis of how 'an error in a few tokens instantly pollutes the context for all other tokens being generated in parallel, initiating a cycle of compounding errors.'

Evidence

Evidence 1 - **Rationale**: Both papers identify the same fundamental problem: the gap between training on clean data and testing with error-prone outputs. The candidate paper explicitly describes this training-inference discrepancy before the original paper's submission. - **Original**: we identify that the fundamental challenge extends beyond error accumulation to a critical discrepancy between the training assumption (seeing clean data) and the test-time autoregressive reality (conditioning on self-generated, error-prone outputs). - **Candidate**: Models are trained exclusively on clean, ground-truth data but are required at inference to generate from their own noisy, intermediate outputs. in a parallel decoding scenario, this discrepancy becomes catastrophic.

Evidence 2 - **Rationale**: Both papers identify that the training assumption of error-free trajectories breaks down during inference when models condition on their own error-prone outputs, leading to error cascades. - **Original**: in training generative dit (fig. 1a), historical trajectories of flow matching are assumed to be error-free. however, this is easily broken in test since the model autoregressively uses previous generations with predictive errors - **Candidate**: existing discrete diffusion models (nie et al., 2025; you et al., 2025; li et al., 2025a) are often plagued by incoherent and hallucinated artifacts (e.g., formatting errors like sequential commas or visually misaligned text) when parallel generation, frequently defaulting to one-token-per-step deco...

Evidence 3 - **Rationale**: Both papers formalize how predictive errors during generation create a trajectory that deviates from the clean training assumption, with the candidate describing the 'cycle of compounding errors' that parallels the original's 'single-clip predictive error.' - **Original**: single-clip predictive error. in eq. 1, the training assumesx t obtained via a clean latentx vid with correct historical trajectory. however, in inference (fig. 2b), this hypothesis is easily broken, since ˜xt is obtained from a predictive trajectory with inherent errors. - **Candidate**: as illustrated in figure 1 (a), an error in a few tokens instantly pollutes the context for all other tokens being generated in parallel, initiating a cycle of compounding errors, which produces a detailed yet entirely fabricated description of the input image.

Evidence 4 - **Rationale**: Both papers identify how errors in previously generated content corrupt the conditioning context for subsequent generation, though the candidate focuses on parallel token generation while the original focuses on cross-clip conditioning. - **Original**: cross-clip conditional error. in fig. 2c, when generating subsequent clips autoregressively, the model useserror-includedframe ˜ximg from ˆxvid (fig. 2b) instead of theclean onex img used in training (eq. 1), leading to a shift in the trajectory start fromx img vid to ˜ximg vid - **Candidate**: however, a significant gap exists between the theoretical promise and the practical reality of these models. existing discrete diffusion models (nie et al., 2025; you et al., 2025; li et al., 2025a) are often plagued by incoherent and hallucinated artifacts (e.g., formatting errors like sequential c...

4. BAgger: Backwards Aggregation for Mitigating Drift in Autoregressive Video Diffusion Models

URL: View paper

Prior Art Analysis

BAgger[39] demonstrates that the training-test hypothesis gap in autoregressive video generation was previously identified and formalized. The candidate paper explicitly defines 'exposure bias' as the fundamental mismatch between training on clean contexts and inference on self-generated frames, which directly corresponds to the original paper's 'training-test hypothesis gap.' Both papers identify the same core problem: models trained on error-free data must operate on error-prone outputs during inference, leading to compounding errors. The candidate's formalization predates the original's claim to be the first to systematically analyze this discrepancy.

Evidence

Evidence 1 - **Rationale**: Both papers identify the identical fundamental problem: the discrepancy between training assumptions (clean/ error-free data) and test-time reality (self-generated, error-prone outputs). BAgger[39] explicitly names this as 'exposure bias' and describes it as a 'mismatch between training on clean contexts and inference on self-generated frames,' which is the same concept as the original's 'training-test hypothesis gap.' - **Original**: we identify that the fundamental challenge extends beyond error accumulation to a critical discrepancy between the training assumption (seeing clean data) and the test-time autoregressive reality (conditioning on self-generated, error-prone outputs). - **Candidate**: autoregressive video models are promising for world modeling via next-frame prediction, but they suffer from exposure bias: a mismatch between training on clean contexts and inference on self-generated frames, causing errors to compound and quality to drift over time.

Evidence 2 - **Rationale**: The original paper's description of error-free training trajectories being broken during test-time autoregressive generation directly matches BAgger[39]'s characterization of exposure bias. Both describe the same phenomenon where training assumes clean inputs but testing uses self-generated, potentially erroneous outputs. - **Original**: in training generative dit (fig. 1a), historical trajectories of flow matching are assumed to be error-free. however, this is easily broken in test since the model autoregressively uses previous generations with predictive errors - **Candidate**: autoregressive video models are promising for world modeling via next-frame prediction, but they suffer from exposure bias: a mismatch between training on clean contexts and inference on self-generated frames, causing errors to compound and quality to drift over time.

Evidence 3 - **Rationale**: BAgger[39] addresses the same training-test gap problem in the context of flow matching and diffusion transformers, demonstrating prior recognition of this fundamental challenge in autoregressive video generation before the original paper's submission. - **Original**: we uncover that the fundamental challenge lies in thehypothesis gap between the training and test. in training generative dit (fig. 1a), historical trajectories of flow matching are assumed to be error-free. however, this is easily broken in test since the model autoregressively uses previous genera... - **Candidate**: bagger trains with standard score or flow matching objectives, avoiding large teachers and long-chain backpropagation through time. we instantiate bagger on causal diffusion transformers and evaluate on text-to-video, video extension, and multi-prompt generation, observing more stable long-horizon m...

5. Ar-diffusion: Asynchronous video generation with auto-regressive diffusion

URL: View paper

Brief Assessment

AR Diffusion[59] addresses training-inference inconsistencies in asynchronous auto-regressive models but focuses on timestep composition constraints and noise scheduling rather than formalizing error types as single-clip predictive error and cross-clip conditional error.

6. End-to-End Training for Autoregressive Video Diffusion via Self-Resampling

URL: View paper

Prior Art Analysis

Self-Resampling[63] demonstrates that the training-test hypothesis gap in autoregressive video generation was previously identified and addressed. The candidate paper explicitly identifies 'exposure bias arising from the train-test mismatch' as a fundamental problem in autoregressive video diffusion models, which directly corresponds to the original paper's concept of the training-test hypothesis gap. Both papers recognize that training assumes clean/error-free data while testing conditions on error-prone outputs. The candidate's 'self-resampling scheme that simulates inference-time model errors on history frames during training' parallels the original's error-recycling approach, indicating prior recognition of this discrepancy.

Evidence

Evidence 1 - **Rationale**: Both papers identify the same fundamental problem: a mismatch between training (clean data) and testing (error-prone outputs) in autoregressive video generation. The candidate uses 'exposure bias' and 'train-test mismatch' terminology to describe what the original calls 'training-test hypothesis gap'. - **Original**: we identify that the fundamental challenge extends beyond error accumulation to a critical discrepancy between the training assumption (seeing clean data) and the test-time autoregressive reality (conditioning on self-generated, error-prone outputs). - **Candidate**: autoregressive video diffusion models hold promise for world simulation but are vulnerable to exposure bias arising from the train-test mismatch.

Evidence 2 - **Rationale**: Both papers recognize that training assumes error-free historical frames while testing uses error-prone ones. The candidate's solution of simulating inference-time errors during training demonstrates prior awareness of this gap. - **Original**: in training generative dit (fig. 1a), historical trajectories of flow matching are assumed to be error-free. however, this is easily broken in test since the model autoregressively uses previous generations with predictive errors - **Candidate**: central to our approach is a self-resampling scheme that simulates inference-time model errors on history frames during training.

7. Recurrent Neural Operators: Stable Long-Term PDE Prediction

URL: View paper

Prior Art Analysis

Recurrent Neural Operators[60] demonstrates that the training-test hypothesis gap and error accumulation in autoregressive prediction were previously identified and formalized in the context of neural operators for PDEs. The candidate paper explicitly addresses the 'mismatch between training and inference' and 'compounding errors' that arise from teacher forcing, providing both theoretical analysis (showing exponential to linear error growth reduction) and a solution framework. This prior work establishes that the fundamental discrepancy between training assumptions and test-time reality was recognized and systematically analyzed before the original paper's submission.

Evidence

Evidence 1 - **Rationale**: Both papers identify the same fundamental problem: a discrepancy between training (clean/ground-truth data) and inference (error-prone autoregressive outputs). The candidate explicitly frames this as a 'mismatch between training and inference' with 'compounding errors', which directly corresponds to the original's 'training assumption versus test-time autoregressive reality' gap. - **Original**: we identify that the fundamental challenge extends beyond error accumulation to a critical discrepancy between the training assumption (seeing clean data) and the test-time autoregressive reality (conditioning on self-generated, error-prone outputs). - **Candidate**: in time-dependent problems, standard training strategies such as teacher forcing introduce a mismatch between training and inference, leading to compounding errors in long-term autoregressive predictions.

Evidence 2 - **Rationale**: The original paper defines 'single-clip predictive error' as arising from training on clean trajectories versus inference on error-prone ones. The candidate addresses this exact issue by proposing recursive training that conditions on the model's own predictions rather than ground-truth, explicitly recognizing and solving the exposure bias problem that creates this error type. - **Original**: single-clip predictive error. in eq. 1, the training assumesx t obtained via a clean latentx vid with correct historical trajectory. however, in inference (fig. 2b), this hypothesis is easily broken, since ˜xt is obtained from a predictive trajectory with inherent errors. - **Candidate**: instead of conditioning each training step on ground-truth inputs, rnos recursively apply the operator to their own predictions over a temporal window, effectively simulating inference-time dynamics during training. this alignment mitigates exposure bias and enhances robustness to error accumulation...

Evidence 3 - **Rationale**: The original paper's 'cross-clip conditional error' describes how errors accumulate when conditioning on error-included outputs. The candidate provides theoretical analysis of this exact phenomenon, characterizing the 'exponential error growth typical of teacher forcing' and demonstrating how to reduce it—indicating prior formal understanding of error accumulation dynamics. - **Original**: cross-clip conditional error. in fig. 2c, when generating subsequent clips autoregressively, the model useserror-includedframe ˜ximg from ˆxvid (fig. 2b) instead of theclean onex img used in training (eq. 1), leading to a shift in the trajectory start - **Candidate**: theoretically, we show that recurrent training can reduce the worst-case exponential error growth typical of teacher forcing to linear growth.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Stable Video Infinity: Infinite-Length Video Generation with Error Recycling View paper
- [1] MimicMotion: High-Quality Human Motion Video Generation with Confidence-aware Pose Guidance View paper
- [2] MotionStream: Real-Time Video Generation with Interactive Motion Controls View paper
- [3] SkyReels-V2: Infinite-length Film Generative Model View paper
- [4] Streamingt2v: Consistent, dynamic, and extendable long video generation from text View paper
- [5] LaVie: High-Quality Video Generation with Cascaded Latent Diffusion Models View paper
- [6] Live Avatar: Streaming Real-time Audio-Driven Avatar Generation with Infinite Length View paper
- [7] Vidgen: Long-Form Text-to-Video Generation with Temporal, Narrative and Visual Consistency for High Quality Story-Visualisation Tasks View paper
- [8] Longlive: Real-time interactive long video generation View paper
- [9] 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors View paper
- [10] Macro-from-micro planning for high-quality and parallelized autoregressive long video generation View paper
- [11] Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models View paper

- [12] Generating long videos of dynamic scenes View paper
- [13] DeepVerse: 4D Autoregressive Video Generation as a World Model View paper
- [14] Video is worth a thousand images: Exploring the latest trends in long video generation View paper
- [15] Tora: Trajectory-oriented Diffusion Transformer for Video Generation View paper
- [16] FreeLong: Training-Free Long Video Generation with SpectralBlend Temporal Attention View paper
- [17] DriveGen3D: Boosting Feed-Forward Driving Scene Generation with Efficient Video Diffusion View paper
- [18] RoboEnvision: A Long-Horizon Video Generation Model for Multi-Task Robot Manipulation View paper
- [19] MALT Diffusion: Memory-Augmented Latent Transformers for Any-Length Video Generation View paper
- [20] SkyReels-Audio: Omni Audio-Conditioned Talking Portraits in Video Diffusion Transformers View paper
- [21] ARLON: Boosting Diffusion Transformers with Autoregressive Models for Long Video Generation View paper
- [22] Adaptive Caching for Faster Video Generation with Diffusion Transformers View paper
- [23] Longvie: Multimodal-guided controllable ultra-long video generation View paper
- [24] Frame In-N-Out: Unbounded Controllable Image-to-Video Generation View paper
- [25] Reward Forcing: Efficient Streaming Video Generation with Rewarded Distribution Matching Distillation View paper
- [26] JoyAvatar: Real-time and Infinite Audio-Driven Avatar Generation with Autoregressive Diffusion View paper
- [27] DreamJourney: Perpetual View Generation with Video Diffusion Models View paper
- [28] Persistent nature: A generative model of unbounded 3d worlds View paper
- [29] StableAvatar: Infinite-Length Audio-Driven Avatar Video Generation View paper
- [30] MagicInfinite: Generating Infinite Talking Videos with Your Words and Voice View paper
- [31] Make-it-4d: Synthesizing a consistent long-term dynamic scene video from a single image View paper
- [32] Compositional Video Synthesis by Temporal Object-Centric Learning View paper
- [33] Self-Forcing++: Towards Minute-Scale High-Quality Video Generation View paper
- [34] LongScape: Advancing Long-Horizon Embodied World Models with Context-Aware MoE View paper
- [35] LMVQ: Label-free Metric-learning for General AI-generated Video Quality Assessment View paper
- [36] STAGE: A Stream-Centric Generative World Model for Long-Horizon Driving-Scene Simulation View paper
- [37] Brick-Diffusion: Generating Long Videos with Brick-to-Wall Denoising View paper
- [38] BlockVid: Block Diffusion for High-Quality and Consistent Minute-Long Video Generation View paper
- [39] BAgger: Backwards Aggregation for Mitigating Drift in Autoregressive Video Diffusion Models View paper
- [40] AI and Generative Models in 360-Degree Video Creation: Building the Future of Virtual Realities View paper
- [41] WorldWeaver: Generating Long-Horizon Video Worlds via Rich Perception View paper
- [42] AI Powered High Quality Text to Video Generation with Enhanced Temporal Consistency View paper
- [43] Markov Decision Process for Video Generation View paper
- [44] TalkVerse: Democratizing Minute-Long Audio-Driven Video Generation View paper
- [45] TempoMaster: Efficient Long Video Generation via Next-Frame-Rate Prediction View paper
- [46] Animate Any Character in Any World View paper
- [47] Rolling Forcing: Autoregressive Long Video Diffusion in Real Time View paper
- [48] Unboxed: Geometrically and Temporally Consistent Video Outpainting View paper
- [49] BulletTime: Decoupled Control of Time and Camera Pose for Video Generation View paper
- [50] Captain Safari: A World Engine View paper
- [51] Structure and content-guided video synthesis with diffusion models View paper
- [52] Dreampose: Fashion image-to-video synthesis via stable diffusion View paper
- [53] Align your latents: High-resolution video synthesis with latent diffusion models View paper
- [54] Frame-Level Captions for Long Video Generation with Complex Multi Scenes View paper
- [55] Make-your-video: Customized video generation using textual and structural guidance View paper
- [56] Storydiffusion: Consistent self-attention for long-range image and video generation View paper
- [57] Dreampose: Fashion video synthesis with stable diffusion View paper
- [58] Text2story: Advancing video storytelling with text guidance View paper
- [59] Ar-diffusion: Asynchronous video generation with auto-regressive diffusion View paper
- [60] Recurrent Neural Operators: Stable Long-Term PDE Prediction View paper
- [61] From Denoising to Refining: A Corrective Framework for Vision-Language Diffusion Model View paper
- [62] Time-series generation by contrastive imitation View paper
- [63] End-to-End Training for Autoregressive Video Diffusion via Self-Resampling View paper
- [64] Self-guided diffusion models View paper
- [65] Cm-gan: Stabilizing gan training with consistency models View paper
- [66] Diffrect: Latent diffusion label rectification for semi-supervised medical image segmentation View paper
- [67] Your diffusion model is secretly a noise classifier and benefits from contrastive training View paper
- [68] ETC: training-free diffusion models acceleration with Error-aware Trend Consistency View paper
- [69] Rethinking Training Dynamics in Scale-wise Autoregressive Generation View paper
- [70] Towards dynamic modeling of visual-vestibular conflict detection View paper
- [71] OViP: Online Vision-Language Preference Learning View paper
- [72] One to Two, Two to All: Towards Multimodal Self-supervised Learning for Earth Observation View paper